Calculating gain for encoded speech transmission by dividing into signal sections and determining weighting factor from periodicity and stationarity

Publication number: DE10026872

Publication date: 2001-10-31

Inventor: FISCHER KYRILL ALEXANDER (DE); ERDMANN CHRISTOPH (DE)

Applicant: DEUTSCHE TELEKOM AG (DE)

Classification:

- international: ***G10L11/02; G10L19/08;*** *G10L11/04;* **G10L11/00; G10L19/00;** (IPC1-7): G10L11/02

- European: G10L11/02; G10L19/08G

Application number: DE20001026872 20000531

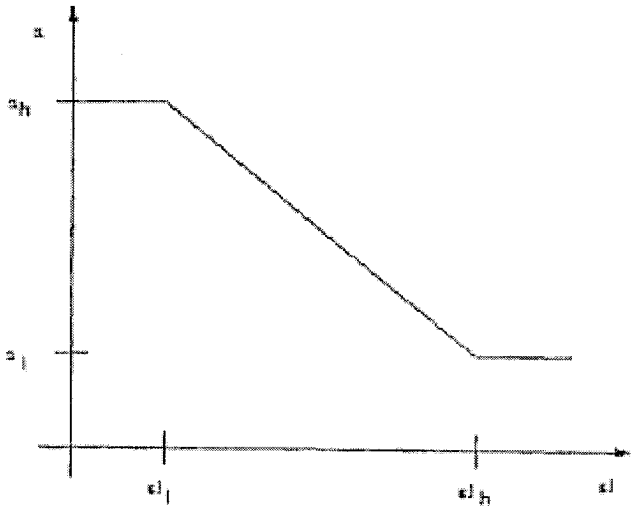
Priority number(s): DE20001026872 20000531; DE20001020863 20000428

Also published as:

- WO0184541 (A1)
- US2003105626 (A1)
- EP1279168 (A0)
- DE10026904 (A1)
- EP1279168 (B1)

Abstract of **DE10026872**

The invention relates to a method for determining voice activity in a signal section of an audio signal. The result, i.e. whether voice activity is present in the section of the signal thus observed, depends upon spectral and temporal stationarity of the signal section and/or prior signal sections. In a first step, the method determines whether there is spectral stationarity in the observed signal section. In a second step, the method determines whether there is temporal stationarity in the signal section in question. The final decision as to the presence of voice activity in the signal section observed depends upon the initial values of both steps.



(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 100 26 872 A 1**

(51) Int. Cl.[7]: G 10 L 11/02

(21) Aktenzeichen: 100 26 872.2
(22) Anmeldetag: 31. 5. 2000
(43) Offenlegungstag: 31. 10. 2001

DE 100 26 872 A 1

(66) Innere Priorität:
100 20 863.0 28. 04. 2000

(71) Anmelder:
Deutsche Telekom AG, 64295 Darmstadt, DE

(72) Erfinder:
Fischer, Kyrill Alexander, Dr., 64347 Griesheim, DE;
Erdmann, Christoph, 52062 Aachen, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht zu ziehende Druckschriften:

DE	197 16 862 A1
DE	40 20 633 A1
DE	694 21 498
DE	694 20 027 T2
DE	690 17 074 T2
US	54 59 814

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

(54) Verfahren zur Berechnung einer Sprachaktivitätsentscheidung (Voice Activity Detector)

(57) Die Erfindung betrifft ein Verfahren zur Bestimmung der Sprachaktivität in einem Signalabschnitt eines Audio-Signals, wobei das Ergebnis, ob Sprachaktivität im betrachteten Signalabschnitt vorliegt sowohl von der spektralen als auch von der zeitlichen Stationarität des Signalabschnitts und/oder von vorangegangenen Signalabschnitten abhängt, wobei das Verfahren in einer ersten Stufe beurteilt, ob im betrachteten Signalabschnitt spektrale Stationarität vorliegt, und daß in einer zweiten Stufe beurteilt wird, ob im betrachteten Signalabschnitt zeitliche Stationarität vorliegt, wobei die endgültige Entscheidung über das Vorhandensein von Sprachaktivität im betrachteten Signalabschnitt von den Ausgangswerten der beiden Stufen abhängig ist.

DE 100 26 872 A 1

Beschreibung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zur Bestimmung der Sprachaktivität in einem Signalabschnitt eines Audio-Signals, wobei das Ergebnis, ob Sprachaktivität im betrachteten Signalabschnitt vorliegt sowohl von der spektralen als auch von der zeitlichen Stationarität des Signalabschnitts und/oder von vorangegangenen Signalabschnitten abhängt.

[0002] Im Bereich der Sprachübertragung und im Bereich der digitalen Signal- und Sprachspeicherung ist die Anwendung spezieller digitaler Codierungsverfahren zu Datenkompressionszwecken weit verbreitet und aufgrund der hohen Datenaufkommen sowie der begrenzten Übertragungskapazitäten zwingend notwendig. Ein für die Übertragung von Sprache besonders geeignetes Verfahren ist das aus der US 4133976 bekannte Code Excited Linear Prediction(CELP)-Verfahren. Bei diesem Verfahren wird das Sprachsignal in kleinen zeitlichen Abschnitten ("Sprachrahmen", "Rahmen", "zeitlicher Ausschnitt", "zeitlicher Abschnitt") von jeweils ca. 5 ms bis 50 ms Länge codiert und übertragen. Jeder dieser zeitlichen Abschnitte bzw. Rahmen wird nicht exakt, sondern nur durch eine Annäherung an die tatsächliche Signalform dargestellt. Die den Signalabschnitt beschreibende Approximation wir dabei im wesentlichen aus drei Komponenten gewonnen, die Decoder-Seitig zur Rekonstruktion des Signals verwendet werden: Erstens einem Filter, das die spektrale Struktur des jeweiligen Signalausschnittes annähernd beschreibt, zweitens einem sog. Anregungssignal, das durch dieses Filter gefiltert wird, und drittens einem Verstärkungsfaktor ("gain"), mit dem das Anregungssignal vor der Filterung multipliziert wird. Der Verstärkungsfaktor ist für die Lautstärke des jeweiligen Abschnitts des rekonstruierten Signals verantwortlich. Das Ergebnis dieser Filterung, stellt dann die Approximation des zu übertragenden Signalstückes dar. Für jeden Abschnitt muß die Information über die Filtereinstellungen und die Information über das zu verwendende Anregungssignal und dessen Skalierung ("gain"), die die Lautstärke beschreibt, übertragen werden. Im allgemeinen werden diese Parameter aus verschiedenen, dem Encoder und Decoder in identischen Kopien vorliegenden Codebüchern gewonnen, so daß zur Rekonstruktion nur die Nummer der am besten geeigneten Codebucheinträge übertragen werden muß. Bei der Codierung eines Sprachsignals sind also für jeden Abschnitt diese am besten geeigneten Codebucheinträge zu bestimmen, wobei alle relevanten Codebuche inträge in allen relevanten Kombinationen durchsucht werden, und diejenigen Einträge ausgewählt werden, die die im Sinne eines sinnvollen Abstandsmaßes kleinste Abweichung zum Originalsignal liefern.

[0003] Es existieren verschiedene Verfahren zur Optimierung der Struktur der Codebücher (z. B. Mehrstufigkeit, Lineare Prädiktion basierend auf den vergangenen Werten, spezifische Abstandsmaße, optimierte Suchverfahren, etc.). Außerdem gibt es verschiedene Verfahren, die den Aufbau und das Durchsuchungsverfahren für die Bestimmung der Anregungsvektoren beschreiben.

[0004] Häufig stellt sich die Aufgabe, den Charakter des im vorliegenden Rahmen befindliche Signales zu klassifizieren, damit die Details der Codierung, z. B. der zu verwendenden Codebücher, etc. bestimmt werden können. Dabei wird häufig auch eine sog. Sprach-Aktivitäts-Entscheidung ("voice activity detection", VAD) getroffen, die angibt, ob der aktuell vorliegende Signalauschnitt ein Sprachsegment oder kein Sprachsegment enthält. Eine solche Entscheidung muss auch bei Anwesenheit von Hintergrundgeräuschen richtig getroffen werden, was die Klassifikation erschwert.

[0005] In dem hier vorgestellten Ansatz wird die Entscheidung der VAD gleichgesetzt mit einer Entscheidung über die Stationarität des aktuellen Signals, so dass also das Ausmaß der Änderung der wesentlichen Signaleigenschaften als Grundlage für die Bestimmung der Stationarität und der damit zusammenhängenden Sprachaktivität verwendet wird. In diesem Sinne ist dann z. B. ein Signalbereich ohne Sprache, der z. B. nur ein gleichbleibend lautes und spektral sich nicht oder nur gering änderndes Hintergrundgeräusch aufweist, als stationär zu bezeichnen. Umgekehrt ist ein Signalausschnitt mit einem Sprachsignal (mit und ohne Anwesenheit des Hintergrundgeräusches) als nicht stationär, also instationär zu bezeichnen. Im Sinne der VAD wird also beim hier vorgestellten Verfahren das Ergebnis "instationär" mit Sprachaktivität gleichgesetzt, während "stationär" bedeutet, dass keine Sprachaktivität vorliegt.

[0006] Da die Stationarität eines Signals keine eindeutig festgelegte Meßgröße ist, wird sie nachfolgend genauer definiert.

[0007] Das vorgestellte Verfahren geht dabei davon aus, dass eine Bestimmung der Stationarität idealerweise von der zeitlichen Änderung des Kurzzeit-Mittelwertes der Energie des Signals ausgehen sollte. Eine solche Schätzung ist aber im allgemeinen nicht direkt möglich, denn sie kann durch verschiedene störende Randbedingungen beeinflußt werden. So hängt die Energie z. B. auch von der absoluten Lautstärke des Sprechers ab, die auf die Entscheidung aber keinen Einfluß haben sollte. Darüber hinaus wird der Energiewert z. B. auch durch das Hintergrundgeräusch beeinflußt. Der Einsatz eines auf einer Energiebetrachtung basierenden Kriteriums ist also nur sinnvoll, wenn der Einfluß dieser möglichen störenden Effekte ausgeschlossen werden kann. Aus diesem Grund ist das Verfahren zweistufig gestaltet: In der ersten Stufe wird bereits eine gültige Entscheidung über die Stationarität getroffen. Falls in der ersten Stufe auf "stationär" entschieden wird, so wird das diesen stationären Signalabschnitt beschreibende Filter neu berechnet und somit an das jeweils letzte stationäre Signal angepaßt. In der zweiten Stufe wird diese Entscheidung jedoch noch einmal nach einem anderen Kriterien getroffen, und damit unter Verwendung der in der ersten Stufe bereitgestellten Werte kontrolliert und gegebenenfalls abgeändert. Diese zweite Stufe arbeitet dabei unter Verwendung eines Energiemaßes. Die zweite Stufe liefert außerdem ein Ergebnis, das von der ersten Stufe bei der Analyse des nachfolgenden Sprachrahmens berücksichtigt wird. Auf diese Weise besteht eine Rückkopplung zwischen diesen beiden Stufen, die sicherstellt, das die von der ersten Stufe gelieferten Werte eine optimale Grundlage für die Entscheidung der zweiten Stufe bilden.

[0008] Die Arbeitsweise der beiden Stufen wird im folgenden einzeln vorgestellt.

[0009] Zunächst wird die erste Stufe vorgestellt, die eine erste Entscheidung basierend auf der Untersuchung der spektralen Stationarität liefert. Betrachtet man das Frequenzspektrum eines Signalabschnitts, so weist es für den betrachteten Zeitraum eine charakteristische Form auf. Ist die Änderung der Frequenzspektren zeitlich aufeinanderfolgender Signalabschnitte hinreichend gering, d. h. die charakteristische Form der jeweiligen Spektren bleibt mehr oder weniger erhalten, so kann man von spektraler Stationarität sprechen.

[0010] Das Ergebnis der Ersten Stufe wird mit STAT1 bezeichnet und das Ergebnis der zweiten Stufe mit STAT2.

STAT2 entspricht auch der endgültigen Entscheidung des hier vorgestellten VAD-Verfahrens. Im folgenden werden Listen mit mehreren Werten in der Form "Listenname[0. .N-1]" beschrieben, wobei über Listenname[k], k = 0. . .N-1 ein einzelner Wert, nämlich der Wert mit dem Index k der Werteliste "Listenname" bezeichnet wird.

Spektrale Stationarität (1. Stufe)

[0011] Diese erste Stufe des Stationaritätsverfahrens erhält als Eingangswerte die folgenden Größen:

- Lineare Prädiktionskoeffizienten des aktuellen Rahmens (LPC_NOW[0. . .ORDER-1]; ORDER = 14)
- ein Maß für die Stimmhaftigkeit des aktuellen Rahmens (STIMM[00. .1])
- Die Anzahl der in der Analyse der zurückliegenden Rahmen durch die zweite Stufe des Algorithmus als "instationär" klassifizierten Rahmen (N_INSTAT2, Werte = 0, 1, 2, usw.)
- verschiedene für die zurückliegenden Rahmen berechnete Werte (STIMM_MEM[0. .1], LPC_STAT1[0. . .ORDER-1])

[0012] Als Ausgangswert liefert die erste Stufe die Werte

- erste Entscheidung über Stationarität: STAT1 (mögliche Werte: "stationär", "instationär")
- Lineare Prädiktionskoeffizienten des letzten als "stationär" klassifizierten Rahmens (LPC_STAT1)

[0013] Die Entscheidung der ersten Stufe basiert primär auf der Betrachtung der sog. spektralen Distanz ("spektraler Abstand", "spektrale Verzerrung", engl.: "spectral distortion") zwischen dem aktuellen und dem vorangegangenen Rahmen. In die Entscheidung gehen außerdem auch die Werte eines Stimmhaftigkeitsmaßes ein, das für die letzten Rahmen berechnet wurde. Die für die Entscheidung verwendeten Schwellenwerte werden außerdem von der Anzahl der unmittelbar zurückliegenden, in der zweiten Stufe als "stationär" klassifizierten Rahmen (d. h. STAT2 = "stationär") beeinflußt. Die einzelnen Berechnungen werden im folgenden erläutert:

a) Berechnung der spektralen Distanz

[0014] Die Berechnung ergibt sich gemäß:

$$SD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log\left[\frac{1}{\left|A\left(e^{j\omega}\right)\right|^2}\right]-10\log\left[\frac{1}{\left|A'\left(e^{j\omega}\right)\right|^2}\right]\right)^2 d\omega}\ .$$

[0015] Dabei bezeichnet

$$10\log\left[\frac{1}{\left|A\left(e^{j\omega}\right)\right|^2}\right]$$

den logarithmierten Einhüllendenfrequenzgang des aktuellen Signalabschnitts, der aus LPC_NOW berechnet wird.

$$10\log\left[\frac{1}{\left|A'\left(e^{j\omega}\right)\right|^2}\right]$$

bezeichnet den logarithmierten Einhüllendenfrequenzgang des vorangegangenen Signalabschnitts, der aus LPC_STAT1 berechnet wird.

[0016] Der Wert von SD wird nach der Berechnung nach unten auf einen Minimalwert von 1.6 begrenzt. Der so begrenzte Wert wird dann als aktueller Wert in eine Liste der vergangenen Werte SD_MEM[0. .9] gespeichert, wobei der am längsten zurückliegende Wert zuvor aus der Liste entfernt wurde.

[0017] Neben dem aktuellen Wert für SD wird auch ein Mittelwert der vergangenen 10 Werte von SD berechnet, der in SD_MEAN gespeichert wird, wobei zur Berechnung die Werte aus SD_MEM verwendet werden.

b) Berechnung der mittleren Stimmhaftigkeit

[0018] Als Eingangswert in die erste Stufe wurden auch die Ergebnisse eines Stimmhaftigkeitsmasses (STIMM[0. .1]) bereitgestellt. (Diese Werte liegen zwischen 0 und 1 und wurden zuvor nach

$$\chi = \frac{\sum_{i=0}^{L-1} s(i) \cdot s(i-\tau)}{\sqrt{\sum_{i=0}^{L-1} s^2(i) \cdot \sum_{i=0}^{L-1} s^2(i-\tau)}}$$

berechnet. Durch Bildung des kurzzeitigen Mittelwertes von χ über den letzten 10 Signalabschnitten (m_{cur}: Index des momentanen Signalabschnitts) folgen die Werte:

$$STIMM[k] = \frac{1}{10} \sum_{i=m_{cur}-10}^{m_{cur}} \chi_i \quad , \quad k=0,\ 1$$

wobei für jeden Rahmen zwei Werte berechnet werden:
STIMM[0] für die erste Rahmenhälfte, und STIMM[1] für die zweite Rahmenhälfte. Hat STIMM[k] einen Wert nahe 0, so ist das Signal eindeutig stimmlos, während ein Wert nahe 1 einen eindeutig stimmhaften Sprachbereich charakterisiert.)

[0019] Um zunächst Störungen im Sonderfall sehr leiser Signale (z. B. vor Signalbeginn) auszuschließen, werden die daraus resultierenden sehr kleinen Werte von STIMM[k] auf 0.5 gesetzt, nämlich dann, wenn ihr Wert zuvor unter 0.05 lag (für k = 0, 1).

[0020] Die so begrenzten Werte werden dann als aktuellste Werte an der Stelle 19 in eine Liste der vergangenen Werte STIMM_MEM[0..19] gespeichert, wobei die am längsten zurückliegenden Werte zuvor aus der Liste entfernt wurden.

[0021] Über die zurückliegenden 10 Werte von STIMM_MEM[ ] wird nun gemittelt, und das Ergebnis wird in STIMM_MEAN abgelegt.

[0022] Die letzten vier Werte von STIMM_MEM, nämlich die Werte STIMM_MEM[16] bis STIMM_MEM[19] werden noch einmal gemittelt und in STIMM4 gespeichert.

c) Berücksichtigung der Anzahl eventuell vorliegender vereinzelter "stimmhaft"-Rahmen

[0023] Sollten bei der Analyse der zurückliegenden Rahmen vereinzelt instationäre Rahmen aufgetreten sein, so wird dies anhand des Wertes von N_INSTAT2 erkannt. In diesem Fall liegt ein Übergang in den "stationär"-Zustand nur einige wenige Rahmen zurück. Die für die zweite Stufe notwendigen LPC_STAT1[ ]-Werte, die in der ersten Stufe bereitgestellt werden, sollen in diesem Übergangsbereich aber noch nicht sofort, sondern erst nach einigen abzuwartenden "Sicherheitsrahmen" auf einen neuen Wert gebracht werden. Aus diesem Grund wird für den Fall, dass N_INSTAT2 > 0 ist, der interne Schwellwert TRES_SD_MEAN, der für die nachfolgende Entscheidung verwendet wird, auf einen anderen Wert gesetzt als sonst:

TRES_SD_MEAN = 4.0 (wenn N_INSTAT2 > 0)

TRES_SD_MEAN = 2.6 (sonst)

d) Entscheidung

[0024] Zur Entscheidung wird zunächst sowohl SD selbst als auch sein kurzzeitlicher Mittelwert über den letzten 10 Signalabschnitten SD_MEAN betrachtet. Liegen beide Maße SD und SD_MEAN unterhalb eines für sie spezifischen Schwellwertes TRES_SD bzw. TRES_SD_MEAN, so wird spektrale Stationarität angenommen.

[0025] Konkret gilt für die Schwellenwerte:

TRES_SD = 2.6 dB

TRES_SD_MEAN = 2.6 oder 4.0 dB (vgl. c)

und es wird entschieden

STAT1 = "stationär" wenn
(SD < TRES_SD) UND (SD_MEAN < TRES_SD_MEAN),
STAT1 = "instationär" (sonst).

[0026] Innerhalb eines Sprachsignales, das gemäß der Zielsetzung der VAD als "instationär" klassifiziert werden sollte, können allerdings kurzzeitig auch Abschnitte auftreten, die nach obigem Kriterium als "stationär" betrachtet werden. Solche Abschnitte können allerdings dann über das Stimmhaftigkeitsmass STIMM_MEAN erkannt und ausgeschlossen werden: Falls der aktuelle Rahmen nach obiger Regel als "stationär" klassifiziert wurde, so kann nach folgender Regel eine Korrektur erfolgen:

STAT1 = "instationär" wenn
(STIMM_MEAN ≥ 0.7) UND (STIMM4 < = 0.56)
oder (STIMM_MEAN < 0.3) UND (STIMM4 < = 0.56)

oder STIMM_MEM[19] > 1.5.

[0027] Damit liegt das Ergebnis der ersten Stufe vor.

e) Vorbereiten der Werte für die zweite Stufe

[0028] Die zweite Stufe arbeitet unter Verwendung einer in dieser Stufe vorbereiteten Liste von Linearen-Prädiktionskoeffizienten, die das zuletzt von dieser Stufe als "stationär" klassifizierte Signalstück beschreiben. In diesem Fall wird LPC_STAT1 durch das aktuelle LPC_NOW überschrieben (update):

LPC_STAT1[k] = LPC_NOW[k], k = 0. . .ORDER-1 wenn STAT1 "stationär"

[0029] Anderenfalls werden die Werte in LPC_STAT1[ ] nicht geändert und beschreiben somit weiterhin den letzten von der ersten Stufe als "stationär" klassifizierten Signalausschnitt.

Zeitliche Stationarität (2. Stufe)

[0030] Betrachtet man einen Signalabschnitt im Zeitbereich, so weist es einen für den betrachteten Zeitraum charakteristischen Amplituden- bzw. Energieverlauf auf. Bleibt die Energie zeitlich aufeinanderfolgender Signalabschnitte konstant, bzw. die Abweichung der Energie ist auf ein hinreichend kleines Toleranzintervall begrenzt, so kann man von zeitlicher Stationarität sprechen. Das Vorliegen einer zeitlichen Stationarität wird in der zweiten Stufe analysiert.

[0031] Als Eingangsgrößen verwendet die zweite Stufe die Werte

- das aktuelle Sprachsignal in abgetasteter Form (SIGNAL[0. . .FRAME_LEN-1], FRAME_LEN = 240)
- VAD-Entscheidung der ersten Stufe: STAT1 (mögliche Werte: "stationär", "instationär")
- die linearen Prädiktionskoeffizienten, die den letzten "stationären" Rahmen beschrieben (LPC_STAT1[0. .13])
- die Energie des Residualsignales des vorherigen stationären Rahmens (E_RES_REF)
- Eine Variable ANFANG, die einen Neubeginn der Werteanpassung steuert (ANFANG, Werte = "true", "false")

[0032] Als Ausgangswert liefert die zweite Stufe die Werte

- abschliessende Entscheidung über Stationarität: STAT2 (mögliche Werte: "stationär", "instationär")
- Die Anzahl der in der Analyse der zurückliegenden Rahmen durch die zweite Stufe des Algorithmus als "instationär" klassifizierten Rahmen (N_INSTAT2, Werte = 0, 1, 2, usw.) und die Anzahl der unmittelbar zurückliegenden stationären Rahmen N_STAT2 (Werte = 0, 1, 2, usw.).
- Die Variable ANFANG, die ggf. auf einen neuen Wert gesetzt wurde.

[0033] Zur VAD-Entscheidung der zweiten Stufe wird die zeitliche Änderung der Energie des Residualsignales verwendet, das mit dem an den letzten stationären Signalabschnitt angepassten LPC-Filters LPC_STAT1[ ] und dem aktuellen Eingangssignal SIGNAL[ ] berechnet wurde. Dabei gehen sowohl eine Schätzung der zuletzt vorliegenden Restsignalenergie E_RES_REF als unterer Referenzwert und ein vorher ausgewählter Toleranzwert E_TOL in die Entscheidung ein. Der aktuelle Restsignal-Energiewert darf dann um nicht mehr als E_TOL über dem Referenzwert E_RES_REF liegen, wenn das Signal als "stationär" gelten soll.

[0034] Die Bestimmung der relevanten Grössen wird im folgenden dargestellt.

a) Berechnung der Energie des Residualsignals

[0035] Das Eingangssignal SIGNAL[0. . .FRAME_LEN-1] des aktuellen Rahmens wird unter Verwendung der in LPC_STAT1[0. .ORDER-1] gespeicherten Linearen Prädiktionskoeffizienten invers gefiltert. Das Resultat dieser Filterung wird als "Residualsignal" bezeichnet und in SPEECH_RES[0. .FRAME_LEN-1] gespeichert.

[0036] Darauf wird die Energie E_RES dieses Residualsignals SIGNAL_RES[ ] berechnet:

E_RES = Summe [SIGNAL_RES[k] · SIGNAL_RES[k]/FRAME_LEN],
k = 0. . .FRAME_LEN-1

und dann logarithmisch dargestellt:

E_RES = 10 · log (E_RES/E_MAX),

wobei

E_MAX = SIGNAL_MAX · SIGNAL_MAX

[0037] SIGNAL_MAX beschreibt den maximal möglichen Amplitudenwert eines einzelnen Abtastwertes. Dieser Wert ist abhängig von der Implementierungsumgebung; in dem der Erfindung zugrundeliegenden Prototyp betrug er beispielsweise

SIGNAL_MAX = 32767;

in anderen Anwendungsfällen ist gegebenenfalls z. B.

SIGNAL_MAX = 1.0

zu setzen.

[0038] Der so berechnete Wert E_RES ist in dB bezüglich des Maximalwertes ausgedrückt. Er liegt somit stets unterhalb von 0, typische Werte betragen etwa –100 dB für Signale mit sehr niedriger Energie und etwa –30 dB für Signale mit vergleichsweise hoher Energie.

[0039] Falls der berechnete Wert E_RES sehr klein ist, so liegt ein Anfangszustand vor, und der Wert von E_RES wird nach unten begrenzt:

wenn (E_RES < –200):
E_RES = –200
ANFANG = true

[0040] Diese Bedingung ist effektiv nur zu Beginn des Algorithmus oder bei sehr langen, sehr ruhigen Pausen erfüllbar, so dass nur zu Beginn der Wert ANFANG = true gesetzt werden kann.

[0041] Der Wert von ANFANG wird unter dieser Bedingung auf false gesetzt:

wenn (N_INSTAT2 > 4):
ANFANG = false

[0042] Um die Berechnung der Referenz-Restsignalenergie auch für den Fall niedriger Signalenergie sicherzustellen, wird folgende Bedingung eingeführt:

wenn (ANFANG = false) UND (E_RES < –65.0):
STAT1 = "stationär"

[0043] Damit wird die Bedingung für die Anpassung von E_RES_REF auch für sehr ruhige Signalpausen erzwungen.

[0044] Durch die Verwendung der Energie des Residualsignales wird implizit eine Anpassung an die zuletzt als stationär klassifizierte Spektralform vorgenommen. Sollte sich das aktuelle Signal gegenüber dieser Spektralform geändert haben, so wird das Residualsignal eine messbar höhere Energie besitzen als in dem Fall eines ungeänderten, gleichmässig fortgesetzten Signals.

b) Berechnung der Referenz-Restsignalenergie E_RES_REF

[0045] Neben dem durch LPC_STAT1[ ] beschriebenen Einhüllendenfrequenzgang des zuletzt von der ersten Stufe als "stationär" klassifizierten Rahmens wird in der zweiten Stufe auch die Residualenergie dieses Rahmens gespeichert und als Referenzwert verwendet. Dieser Wert wird mit E_RES_REF bezeichnet. Sie wird hier immer genau dann neu festgesetzt, wenn die erste Stufe den aktuellen Rahmen als "stationär" klassifiziert hat. In diesem Fall wird als neuer Giert für diese Referenzenergie E_RES_REF der zuvor berechnete Wert E_RES verwendet:

[0046] Wenn STAT1 = "stationär" dann setze

E_RES_REF = E_RES wenn
(E_RES < E_RES_REF + 12 dB) ODER
(E_RES_REF < –200 dB) ODER
(E_RES < –65 dB)

[0047] Die erste Bedingung beschreibt den Normalfall: Eine Anpassung von E_RES_REF findet somit fast immer statt, wenn STAT1 = "stationär" ist, denn der Toleranzwert von 12 dB ist bewußt grosszügig gewählt. Die anderen Bedingungen sind Spezialfälle; sie sorgen für eine Anpassung zu Beginn des Algorithmus und für eine Neuschätzung bei sehr niedrigen Eingangswerten, die in jedem Falle als neuer Referenzwert für stationäre Signalabschnitte gelten sollen.

c) Bestimmung des Toleranzwertes E_TOL

[0048] Der Toleranzwert E_TOL gibt für das Entscheidungskriterium eine maximale erlaubte Änderung der Energie des Residialsignales gegenüber derjenigen der vorherigen Rahmens an, damit der aktuelle Rahmen als "stationär" gelten kann. Zunächst wird gesetzt

E_TOL = 12 dB

[0049] Dieser vorläufige Wert wird nachfolgend jedoch unter bestimmten Bedingungen korrigiert:

```
wenn N_STAT2 <= 10:
     E_TOL = 3.0

sonst
     wenn E_RES < -60:
          E_TOL = 13.0
     sonst
          wenn E_RES > -40:
               E_TOL = 1.5
          sonst
               E_TOL = 6.5
```

[0050] Mit der ersten Bedingung wird sichergestellt, dass eine bisher nur kurz bestehende Stationarität sehr leicht verlassen werden kann, indem durch die niedrige Toleranz E_TOL leichter auf "instationär" entschieden wird. Die anderen Fälle beinhalten Anpassungen, die für verschiedene Spezialfälle jeweils günstigste Werte vorsehen (Abschnitte mit sehr niedriger Energie sollen schwerer als "instationär" klassifiziert werden, Abschnitte mit vergleichsweise hoher Energie sollen leichter als "instationär" klassifiziert werden).

d) Entscheidung

[0051] Die eigentliche Entscheidung findet nun unter Verwendung der zuvor berechneten und angepassten Werte E_RES, E_RES_REF und E_TOL statt. Ausserdem wird sowohl die Anzahl aufeinanderfolgender "stationärer" Rahmen N_STAT2 als auch die Anzahl zurückliegender instationärer Rahmen N_INSTAT2 auf aktuelle Werte gesetzt.

[0052] Die Entscheidung erfolgt nach:

```
wenn ( E_RES > E_RES_REF + E_TOL):
     STAT2       = "instationär"
     N_STAT2     = 0
     N_INSTAT2   = N_INSTAT2 + 1
sonst
     STAT2       = "stationär"
     N_STAT2     = N_STAT2 + 1
     wenn N_STAT2 > 16:
          N_INSTAT = 0
```

[0053] Der Zähler der zurückliegenden stationären Rahmen N_STAT2 wird also sofort beim Auftreten eines instationären Rahmens auf 0 gesetzt, während der Zähler für die zurückliegenden instationären Rahmen N_INSTAT2 erst nach dem Vorliegen einer bestimmten Anzahl (im realisierten Prototyp: 16) von aufeinanderfolgenden stationären Rahmen auf 0 gesetzt wird. N_INSTAT2 wird als Eingangswert der ersten Stufe verwendet, und hat dort Einfluß auf die Entscheidung der ersten Stufe. Konkret wird über N_INSTAT2 verhindert, dass die erste Stufe den das Einhüllendenspektrum beschreibenden Koeffizientensatz LPC_STAT1[ ] neu bestimmt, bevor gesichert ist, dass tatsächlich ein neuer stationärer Signalabschnitt vorliegt. Kurzzeitige oder vereinzelte STAT2 = "stationär"-Entscheidungen können also auftreten, aber erst nach einer bestimmten Anzahl aufeinanderfolgender als "stationär" klassifizierter Rahmen wird auch der das Einhüllendenspektrum beschreibenden Koeffizientensatz LPC_STAT1[ ] für den dann vorliegenden stationären Signalabschnitt in der ersten Stufe neu bestimmt.

[0054] Entsprechend der für die zweite Stufe vorgestellten Arbeitsweise und der vorgestellten Parameter wird die zweite Stufe eine STAT1 = "stationär"-Entscheidung der ersten Stufe niemals zu "instationär" abändern, sondern wird in diesem Falle immer ebenfalls auf STAT2 = "stationär" entscheiden.

[0055] Eine "STAT1 = "instationär"-Entscheidung der ersten Stufe kann dagegen von der zweiten Stufe zu einer STAT2 = "stationär"-Entscheidung korrigiert werden, oder auch als STAT2 = "instationär" bestätigt werden. Dies ist insbesondere dann der Fall, wenn die spektrale Instationarität, die in der ersten Stufe zu STAT1 = "instationär" geführt hat, lediglich durch vereinzelte spektrale Schwankungen des Hintergrundsignales verursacht wurde. Dieser Fall wird jedoch

in der zweiten Stufe unter Berücksichtigung der Energie neu entschieden.

[0056] Es versteht sich von selbst, daß die Algorithmen zur Bestimmung der Sprachaktivität, der Stationarität und der Periodizität den jeweils gegebenen Umständen entsprechend angepaßt werden müssen bzw. können. Die einzelnen o. a. Schwellwerte und Funktionen sind lediglich exemplarisch und müssen in der Regel durch eigene Versuche herausgefunden werden.

Patentansprüche

1. Verfahren zur Bestimmung der Sprachaktivität in einem Signalabschnitt eines Audio-Signals, wobei das Ergebnis, ob Sprachaktivität im betrachteten Signalabschnitt vorliegt sowohl von der spektralen als auch von der zeitlichen Stationarität des Signalabschnitts und/oder von vorangegangenen Signalabschnitten abhängt, **dadurch gekennzeichnet**, daß das Verfahren in einer ersten Stufe beurteilt, ob im betrachteten Signalabschnitt spektrale Stationarität vorliegt, und daß in einer zweiten Stufe beurteilt wird, ob im betrachteten Signalabschnitt zeitliche Stationarität vorliegt, wobei die endgültige Entscheidung über das Vorhandensein von Sprachaktivität im betrachteten Signalabschnitt von den Ausgangswerten der beiden Stufen abhängig ist.
2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß zur Ermittlung der spektralen Stationarität sowie der Energieveränderung (zeitliche Stationarität) mindestens ein zeitlich vorangegangener Signalabschnitt berücksichtigt wird.
3. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß jeder Signalabschnitt in mindestens zwei Unterabschnitte aufgeteilt wird, die sich überlappen können, wobei für jeden Unterabschnitt die Sprachaktivität bestimmt wird.
4. Verfahren nach Anspruch 3, dadurch gekennzeichnet, daß für die Beurteilung der Sprachaktivität eines zeitlich nachfolgenden Signalabschnitts die ermittelten Werte für die Sprachaktivität der einzelnen Unterabschnitte jedes vorangegangen Signalabschnitts berücksichtigt werden.
5. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß in der ersten Stufe die spektrale Verzerrung (engl. spectral distortion) zwischen dem aktuell betrachteten Signalabschnitt und dem oder den vorangegangenen Signalabschnitten ermittelt wird.
6. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die erste Stufe eine erste Entscheidung über die Stationarität des betrachteten Signalabschnitts trifft, wobei eine Ausgangsgröße STAT1 die Werte "stationär" oder "instationär" annehmen kann.
7. Verfahren nach Anspruch 6, dadurch gekennzeichnet, daß die Entscheidung über die Stationarität auf Basis der zuvor ermittelten linearen Prädikationskoeffizienten des aktuellen Signalabschnitts LPC_NOW[ ] und einem zuvor ermittelten Maß für die Stimmhaftigkeit des betrachteten Signalabschnitts erfolgt.
8. Verfahren nach Anspruch 7, dadurch gekennzeichnet, daß zusätzlich die Anzahl der in der Analyse der zurückliegenden Signalabschnitte durch die zweite Stufe als "instationär" klassifizierten Signalabschnitte N_INSTAT2 für die Bewertung von STAT1 berücksichtigt werden.
9. Verfahren nach Anspruch 7 oder 8, dadurch gekennzeichnet, daß zusätzlich für die zurückliegenden Rahmen berechnete Werte wie z. B. STIMM_MEM[0. .1], LPC_STAT1[ ] bei der Berechnung eines Wertes für STAT1 berücksichtigt werden.
10. Verfahren nach einem der vorherigen Ansprüche, dadurch gekennzeichnet, daß die erste Stufe zusätzlich zu dem Ausgangswert STAT1 einen weiteren Ausgangswert LPC_STAT1[ ] liefert, der von LPC_NOW[ ] und STAT1 abhängig ist.
11. Verfahren nach einem der vorherigen Ansprüche, dadurch gekennzeichnet, daß in der zweiten Stufe zur Beurteilung, ob zeitliche Stationarität vorliegt, zumindest folgende Eingangsgrößen verwendet werden:
    - Signalabschnitt in abgetasteter Form;
    - STAT1 (Entscheidung der ersten Stufe).
12. Verfahren nach Anspruch 11, dadurch gekennzeichnet, daß zusätzlich folgende Eingangsgrößen in der zweiten Stufe verwendet werden:
    - die linearen Prädikationskoeffizienten LPC_STAT1[ ], die den letzten stationären Signalabschnitt beschreiben;
    - die Energie E_RES_REF des Residualsignales des vorherigen stationären Signalabschnitts;
    - eine Variable ANFANG, die einen Neubeginn der Werteanpassung steuert, wobei die Variable ANFANG die Werte "wahr" und "falsch" annehmen kann.
13. Verfahren nach einem der vorherigen Ansprüche, dadurch gekennzeichnet, daß immer wenn STAT1 gleich "stationär" die zweite Stufe als Ergebnis für STAT2 "stationär" ausgibt.
14. Verfahren nach einem der vorherigen Ansprüche, dadurch gekennzeichnet, daß der Wert von STAT2 das Maß für die Sprachaktivität des betrachteten Signalabschnitts ist.